~~# No cake # No chhr~~ # No misty # No white suger. # No cold drinks.

CT # 01

Basic data mining.
KDD process.
Data-pre Processing.
Data Normalization.
— 1 set.

Classification ⟶ 1 set.

****

****

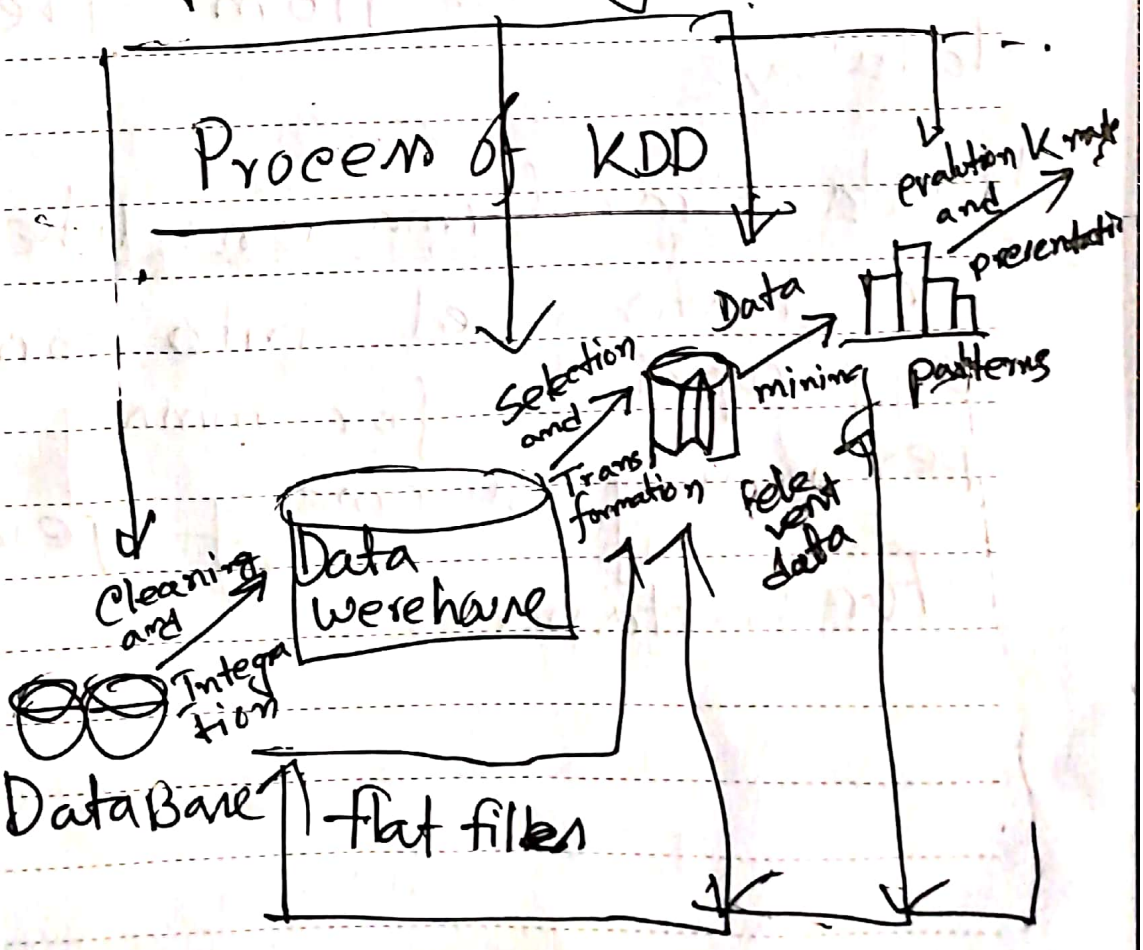
Data warehouse : Data warehouse is a subject oriented, Integrated, Non-volatile, Time varient collection of data in a support of management's decisions.

Date ..../..../....

KDD : Knowledge discovery :

KDD in the process of finding useful information and pattern in data.

Data mining: Data mining is the use of algorithm to extract the information and pattern derived by the Knowledge discovery (KDD) process.



1. Data cleaning: to remove noise and inconsistent data.

2. Data integration: where multiple data soureen may be combined..

3. Data selection: where data relevent to the analysis tark are retrieved from the databare.

4. data transformation: where data are transformed into forms appropriate for mining by performing summary operations, for instance:-

189 | 5 | 76068 | 12 | 13 | 12

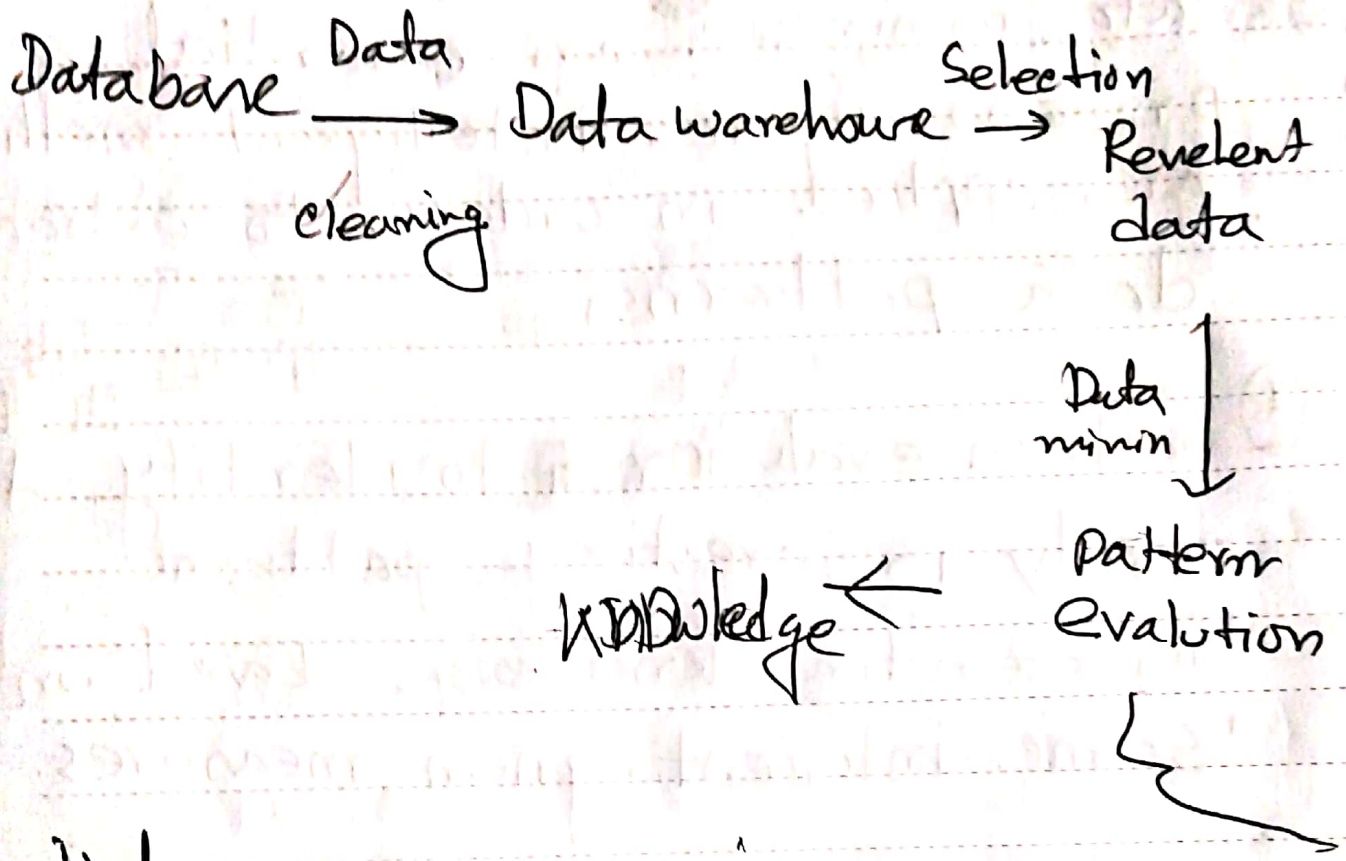
Date ..../..../....

(5) Data mining: an essential process where intelligent methods are applied in order to extract data patterns.

(6) pattern evalution: to identify the truly interesting patterns representing knowledge based on some interestingness measures.

(7) Knowledge presentation: where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

Date ..../..../....



Data preprocessing:

Data pre-processing is a data mining technique that involves transforming raw data into an understandable format. It is a proven method of solving such incomplete data, inconsistent or lacking in certain behaviors and in

Date ..../..../....

likely to contain many errors

steps during pre-processing:

1. Data cleaning: Data in cleansed through processes such as filling in missing values, smoothing the noisy data, resolving the inconsistencies in the data.

2. Data integration: Data with different representations are put together and conflicts within the data are resolved.

* Data transformation: Data is normalized, aggregated and generalized.

Date ...../...../.....

Data Reduction: This step aims to present a reduced representation of the data in a data warehouse.

* Data Discretization: Involves the reduction of a number of values of a continuous attribute by dividing the range of attribute intervals.

Date ...../...../.....

Data Normalization: Normalization is used to scale the data of an attribute so that it falls in a smaller range, such as -1.0 to 1.0 or 0.0 to 1.0. It is generally useful for classification algorithms.

Methods of Data Normalization:

① Decimal scaling: It normalizes by moving the decimal point of values of the data. To normalize the data by this technique, we divide each value of the data by the maximum absolute value of data.

$$V_i' = \frac{V_i}{10^j}$$

where j is the smallest integer such that $\max(|v_i'|) < 1$.

Min-Max Normalization: Linear transformation is performed on the original data. Minimum and maximum value from data is fetched and each value is replaced according to the following formula

Date ..../..../....

$$V' = \frac{V - min(A)}{max(A) - min(A)} \times ((new\_max(A) - new\_min A)) + new\_min(A).$$

Z-score normalization:

In this technique, values are normalized based on mean and standard deviation of the data $\bar{A}$.

The formula used is:

$$v' = \frac{V - \bar{A}}{\sigma_A}$$

A = mean, $\bar{A}$ = standard deviation

Date / /

Where use normalization:

(i) Data reduction.

(ii) Data transformation.

(i) It is used to handle huge amount of data. while working with huge volume of data, analysis became harder in such cases. In order to get rid of this, we uses data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis cost.

(i) Data cube Aggregation.

(ii) Attribute subset selection.

(iii) Numerosity Reduction.

(iv) Dimensionality Reduction.

(v) Data compression.

Date ...../...../.....

Data transformation is the process of converting data or information from one format to another.

① Data mapping.

② Code generation.